Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests

Pulm Function Study Investigators

Published in:
European Respiratory Journal

DOI:
10.1183/13993003.01660-2018

Publication date:
2019

ERJ Express. Published on February 14, 2019 as doi: 10.1183/13993003.01660-2018

EUROPEAN RESPIRATORY *journal*

FLAGSHIP SCIENTIFIC JOURNAL OF ERS

Early View

Original article

Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests

Marko Topalovic, Nilakash Das, Pierre- Régis Burgel, Marc Daenen, Eric Derom, Christel Haenebalcke, Rob Janssen, Huib A. M. Kerstjens, Giuseppe Liistro, Renaud Louis, Vincent Ninane, Christophe Pison, Marc Schlesser, Piet Vercauter, Claus F. Vogelmeier, Emiel Wouters, Jokke Wynants, Wim Janssens

Please cite this article as: Topalovic M, Das N, Burgel P-R, *et al*. Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests. *Eur Respir J* 2019; in press (https://doi.org/10.1183/13993003.01660-2018).

Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests

Marko Topalovic[1], Nilakash Das[1], Pierre- Régis Burgel[2], Marc Daenen[3], Eric Derom[4], Christel Haenebalcke[5], Rob Janssen[6], Huib A.M. Kerstjens[7], Giuseppe Liistro[8], Renaud Louis[9], Vincent Ninane[10], Christophe Pison[11], Marc Schlesser[12], Piet Vercauter[13], Claus F. Vogelmeier[14], Emiel Wouters[15], Jokke Wynants[16], and Wim Janssens[1], on behalf of the Pulmonary Function Study Investigators*.

[1]*Respiratory Medicine, University Hospital Leuven, Chronic Diseases, Metabolism and Ageing, KU Leuven, Belgium;*

[2]*Cochin Hospital, Assistance Publique Hôpitaux de Paris, Université Paris Descartes, Sorbonne Paris Cité, Paris, France;*

[3]*Department of Respiratory Medicine, Hospital Oost-Limburg, Genk, Belgium;*

[4]*Department of Respiratory Medicine, Ghent University Hospital, Ghent, Belgium;*

[5]*Department of Respiratory Medicine, AZ Sint-Jan Hospital, Bruges, Belgium;*

[6]*Department of Pulmonary Medicine, Canisius Wilhelmina Hospital, Nijmegen, The Netherlands;*

[7]*Department of Pulmonary Medicine and Tuberculosis, University of Groningen, and University Medical Center Groningen, Groningen, The Netherlands;*

[8]*Université Catholique de Louvain (UCL), Department of Pneumology, Cliniques universitaires St-Luc, Brussels, Belgium;*

[9]*Department of Respiratory Medicine, University Hospital, Liege, Belgium;*

[10]*Department of Respiratory Medicine, Saint-Pierre Hospital, Université Libre de Bruxelles, Brussels, Belgium;*

[11]*Service hospitalier universitaire de Pneumologie et Physiologie, Centre Hospitalier Universitaire Grenoble Alpes, Université Grenoble Alpes, France;*

[12]*Department of Pulmonary Medicine, Centre Hospitalier de Luxembourg, Luxembourg;*

[13]*Department of Respiratory Medicine, Onze-Lieve-Vrouw Hospital, Aalst, Belgium;*

[14]*Department of Medicine, Pulmonary and Critical Care Medicine, University Medical Center Giessen and Marburg, Marburg, Germany, member of the German Center for Lung Research (DZL);*

[15]*Department of Respiratory Medicine, Maastricht University Medical Center, Maastricht, The Netherlands;*

[16]*Department of Pneumology, Jessa Hospital, Hasselt, Belgium;*

Corresponding and responsible author: Dr Wim Janssens, Respiratory Medicine, University Hospital Leuven, Chronic Diseases, Metabolism and Ageing, KU Leuven, Herestraat 49, 3000 Leuven, Belgium; wim.janssens@uzleuven.be

ABSTRACT

The interpretation of pulmonary function tests (PFTs) to diagnose respiratory diseases is built on expert opinion which relies on the recognition of patterns and clinical context for the detection of specific diseases. In the study, we aimed to explore the accuracy and inter-rater variability of pulmonologists when interpreting PFTs and compared it against that of artificial intelligence (AI)-based software which was developed and validated in more than 1500 historical patient cases.

120 pulmonologists from 16 European hospitals evaluated 50 cases comprising with PFT and clinical information resulting in 6000 independent interpretations. AI software examined the same data. ATS/ERS guidelines were used as the gold standard for PFT pattern interpretation. The gold standard for diagnosis was derived from clinical history, PFT and all additional tests.

The pattern recognition of PFTs by pulmonologists (senior 73%, junior 27%) matched the guidelines in 74.4% (±5.9) of the cases (range: 56-88%). The inter-rater variability of 0.67 (kappa) pointed to a common agreement. Pulmonologists made correct diagnoses in 44.6% (±8.7) of the cases (range: 24-62%) with a large inter-rater variability (kappa= 0.35). The AI-based software perfectly matched the PFT pattern interpretations (100%) and assigned a correct diagnosis in 82% of all cases ($p<0.0001$ for both measures).

The interpretation of PFTs by pulmonologists leads to marked variations and errors. AI-based software provides more accurate interpretations and may serve as a powerful decision support tool to improve clinical practice.

INTRODUCTION

Pulmonary function testing (PFT) is our primary tool to evaluate the function of the respiratory system.[1] In practice, the interpretation is based on expert opinion and involves the recognition of a pattern (obstructive, restrictive, mixed, and normal) and the grading of its severity according to international guidelines.[2-4] To arrive at the final diagnosis the results of PFTs are combined with patient information, symptoms and possibly, the results of other tests, such as imaging, blood analysis, biopsies, and exercise tests.[5, 6]

In 2005, an American Thoracic Society/European Respiratory Society (ATS/ERS) task force designed a simplified algorithm to assess lung function in clinical practice.[2] However, when these recommended guidelines were translated into software for diagnostic decision support, it led to only 38% of correct disease predictions. Adding patient characteristics into such an algorithm improved the accuracy to 68%, highlighting a vast potential for automated diagnostic labelling when combining PFTs with clinical information.[7] In fact, the Belgian pulmonary function study (BPFS) demonstrated that experts panels could reach 77% accuracy when predicting the diagnosis based on PFTs and clinical history alone.[8] Although one may doubt if a computer algorithm carries any added value to a group of experts, the question if it may help individual readers is yet to be answered.

The number of successful applications of artificial intelligence (AI) is quickly rising. Supported by a number of outstanding achievements in the field and because of its unlimited potential to deal with big data, high expectations are also emerging for healthcare. For instance, one study demonstrated the ability of an AI algorithm to identify and classify skin cancer with similar expertise as 21 board-certified dermatologists.[9] Another study reached the same performance when analysing retinal fundus images for the identification of diabetic retinopathy.[10] Moreover, there are multiple examples from radiology in detecting traces of breast and lung cancer.[11, 12] Notwithstanding these technical superiorities of AI-based

systems, translation into clinical practice with broad acceptance has been very challenging.[13-15] Since PFTs are entirely standardised and used worldwide[16], they are ideally suited for the development of AI algorithms for test interpretation and diagnostics. PFTs provide an extensive series of numeric outputs, easily controllable by the computers, yet its patterns are not always easily perceptible or appropriately recognised by the human eye. Moreover, the example of automated interpretation for electrocardiograms which is widely adopted and standardised in most equipment highlights its potential use.

In this study, we hypothesise that AI can improve the clinical reading of PFTs and overcome the variable test interpretation of individual pulmonologists. We explored the accuracy and inter-rater variability of pulmonologists when interpreting patterns of PFTs and when suggesting a specific category of respiratory disease diagnosis based on limited clinical information and PFTs. Secondly, we compared the pulmonologists' performance with that of the AI-based software developed and validated in more than 1500 historical cases.

METHODS

Study design

In this multicentre non-interventional study, 120 pulmonologists from 16 hospitals in 5 European countries participated. They independently evaluated complete PFTs (pre-and/or post-bronchodilator spirometry, whole-body plethysmography for lung volumes and airway resistance, and diffusing capacity) and limited clinical information (smoking history, cough, sputum and dyspnoea) of 50 randomly selected patients, admitted to the University Hospital of Leuven (Belgium) for a respiratory problem. Evaluation sessions were performed in each hospital in the period from August 15th, 2017 till December 13th, 2017. All pulmonologists independently examined different patient cases according to a pre-established protocol by

providing: **A**/ PFT pattern interpretation: obstructive, restrictive, mixed or normal pattern, **B**/ choice of one preferred diagnostic category: 1/asthma, 2/chronic obstructive pulmonary disease (COPD), 3/other obstructive diseases (OBD, including bronchiectasis, bronchiolitis and cystic fibrosis), 4/interstitial lung disease (including idiopathic pulmonary fibrosis, nonspecific interstitial pneumonitis and sarcoidosis), 5/pulmonary vascular disease (including pulmonary hypertension, embolism and vasculitis), 6/neuromuscular disease (including paralysis of the diaphragm, poliomyelitis, myopathy), 7/thoracic deformity (including pneumectomy, lobectomy, chest wall problems, kyphoscoliosis), 8/healthy and 9/other diseases. **C**/ confidence in their decision on a Likert scale: from 1 point ("absolutely not sure") till 5 points ("absolutely sure"). An example is shown in the supplementary material; **S1 – S2**. **D**/ Finally, the same patient files were examined by an in-house developed AI-based software for PFT interpretation and diagnostic suggestion.

Study Population

The study included a random sample of 50 subjects prospectively collected at the outpatient clinic in August 2017. All enrolled subjects were Caucasians older than 18 years who had performed a complete PFT and provided clinical information. The gold standard diagnosis was derived from clinical history, PFT, and all necessary additional tests, and finally confirmed by an expert panel in Leuven. This ad-hoc expert panel consisted of 3 experienced clinicians that reviewed all baseline and clinical follow-up data to agree on a final gold standard diagnosis out of the 9 categories. Consensus was reached for all these cases. Baseline characteristics are shown in **Table 1**, covering a wide range of respiratory diseases that may present with an abnormal PFT. Other conditions (such as lung cancer, cardiovascular disease, ear-nose-throat problem) were excluded from the test sample (n=3). The Ethics Committee of the University Hospital in Leuven approved study protocol (study number S60619, approved on August 4th, 2017). The study design can be found on

www.clinicaltrials.gov (NCT03264417). All included patients provided informed consent for the use of their data (S60243, approved on June 23rd, 2017).

AI Software

The development of software for automated reading of PFTs was performed in R language and its machine learning framework. The software used as input the same lung function data as presented to the pulmonologists (absolute values, percent predicted of normal reference values and z-scores, also shown in **S1**) combined with the patient characteristics age, pack-years, sex, and body mass index (BMI). For pattern interpretations, the PFT algorithm was in line with ATS/ERS strategies.[2] However, the engine for complex diagnostic categorisation had to be developed, and a machine learning approach was adopted.

The machine learning model was built using data from 1430 subjects used in our previous work to ensure a broad variety of data.[7, 8, 17] This data came from two cohorts: 1/ BPFS[8], a prospective cohort study that enrolled a clinical population-based sample (n=851) of all successive undiagnosed patients admitted for the first time to one of the 33 participating Belgian hospitals due to respiratory symptoms; 2/a retrospectively collected PFT data cohort of patients followed at the outpatient clinic of the University Hospital of Leuven based on predefined established diagnoses (neuromuscular disease (n=112), chest-pleural wall problems, including pneumectomy and lobectomy (n=64), pulmonary vascular disease (n=76), other obstructive diseases (n=100), COPD (n=47), asthma(n=40), healthy (n=50), interstitial lung disease (n=90)). Briefly, all subjects were Caucasians between 18 and 85 years old who had performed a complete PFT (including post-bronchodilator spirometry, whole-body plethysmography for lung volumes and airway resistance, and diffusing capacity). The final diagnosis was established with all additional tests deemed necessary by the responsible clinician, the patients' history, and PFTs. Subsequently, it was validated by an ad-hoc installed expert panel (BPFS) or by the clinical expert panel taking care of the patients

in follow-up (Leuven data). The expert discussion of the BPFS were organized during the local meetings of physicians, at which all individual cases were presented to obtain a final diagnosis by consensus. In case there was disagreement, voting was used for a final gold standard diagnosis and if needed, a secondary diagnosis [8]. For the retrospective PFT data collection of patients followed at the University Hospital of Leuven, corresponding medical records were verified on the final diagnosis. For the few cases in which there was doubt on the diagnosis, the PFT data were not extracted and these cases were rejected. Internal 10-fold cross-validation tuned the machine learning model with the best model resulting in the diagnostic accuracy of 74%. To obtain an unbiased estimate of accuracy and validate findings, the model was run at the Leuven pulmonary service on a randomly selected sample of 136 subjects. The model demonstrated a consistent diagnostic accuracy of 76%.[17] Probabilistic output for each of the diagnostic categories obtained by the machine learning model is summarised in the report (Supplement **S3**).

Pulmonary Function Tests

All PFTs were performed with standardised equipment by respiratory technicians (Masterlab, Würzburg, Germany), according to the ATS/ ERS criteria.[18] Spirometry data, as well as plethysmography and single-breath diffusing capacity data, were given as absolute values, but also expressed as percent predicted of normal reference values and as Z-scores.[19-21] In the current prospective study, these data were presented to the AI software and pulmonologist, the latter having also access to the corresponding flow-volume loops, plethysmography and diffusing capacity manoeuvers.

Statistical analysis

Statistical analysis was performed using R software version 3.3.3. (Vienna, Austria: R Foundation for Statistical Computing; 2017). Figures were produced using GraphPad Prism

version 6 (GraphPad Software, La Jolla USA). The inter-observer agreements were assessed using Fleiss' Kappa for multiple raters on categorical data. Interpretative strategies for lung function tests from ATS/ERS task force were used as the gold standard to define a correct lung function pattern.[2] Preferred diagnostic category, by pulmonologists or software, was considered as correct if it corresponded to the gold standard diagnosis made historically by the expert panel based on all data. For both measures, PFT pattern interpretation and diagnostic category suggestion, accuracy was defined as the percentage of correctly labelled cases. The T-test and Mann-Whitney U test were used to evaluate differences between groups with normal and nonparametric distribution, respectively. The Kruskal-Wallis test was used to determine statistical difference between multiple groups. One-sample T-test was used to assess the difference of AI performance and the average accuracy of pulmonologists.

RESULTS

There were 120 pulmonologists who all together made 6000 evaluations of PFTs with clinical information. The pulmonologist group consisted of more senior level members (n=88, established pulmonologists) than junior members (n=32, pulmonologists in training). A minimal number of five pulmonologists per centre was needed to participate.

A/ PFT pattern interpretations

Applying the ATS/ERS interpretative strategies for PFTs revealed that the population consisted of 18 patients with obstructive, 10 with restrictive and 22 with normal lung function pattern, while there were no subjects with a mixed pattern. The interpretations of 118 pulmonologists (data were missing from 2) matched with the reference PFT pattern in 74.4 (±5.9)% of the 50 cases, ranging from 56% to 88% per individual. The identification of a restrictive pattern was more difficult (positive predictive value (PPV)= 59% and sensitivity=

75%) as compared to normal and obstructive patterns (**Table 2**). Even though a mixed pattern was not present, 376 (=6%) cases were interpreted as mixed. A Kappa of 0.67 signified a considerable inter-rater variability or disagreement between different pulmonologists. When the accuracy between different centres was compared, no significant differences in correct detections were found (p=0.06) (**Figure 1A**). There were no significant differences between university and non-university centres (p=0.06) or between senior and junior readers (p=0.49). Interestingly, out of the 285 misclassified normal patterns falsely labelled into an obstructive pattern, 216 (=76%) were on the 4 cases having a FEV_1/FVC ratio above lower limits of normal but still below 0.7 fixed cut-off.

B/ Preferred diagnostic categories

For an individual pulmonologist, it was rather difficult to assign a correct preferred diagnostic category based on complete PFT data and clinical information. The mean accuracy of 6000 evaluations was only 44.6 (±8.7)%, and it ranged from 39% to 51% per centre and from 24% to 62% per individual pulmonologist (**Figure 1B**). A low kappa score of 0.35 was indicative of a common disagreement between pulmonologists. Interestingly, age or clinical experience of the examiners did not influence the mean accuracy (Seniors= 45 (±4.2)% vs Juniors= 43.6 (±4.8)%, p=0.46). Likewise, results were neither different between the hospitals (p=0.44), nor affected by hospital type (university= 44.1 (±9.4)% vs non-university= 45.2 (±7.8)%, p=0.47) or by country (p=0.26).

Due to a higher sensitivity, patterns of healthy subjects (true positive rate= 71%) and subjects with COPD (true positive rate= 65%) were more often identified on lung function than any of the other categories. Patient cases of less prevalent conditions, without a straightforward pattern ("fingerprint") on lung function, were more difficult for the pulmonologists (thoracic deformity and neuromuscular disease, true positive rate= 25%) (asthma, true positive rate= 20%). Detailed statistical group comparison is shown in **Table 3 and Figure S4**.

C/ Confidence in decision-making

Rarely, pulmonologists were "absolutely not sure" (in 2.7% of cases) or "not sure" (11.5%) when suggesting the preferred diagnostic category. Most commonly they were "sure" (36.5%) and "absolutely sure" (16%) in their decisions. Higher confidence in diagnostic suggestion was observed in decisions that were correct ($p<0.0001$) as compared to the incorrect decisions. However, high confidence did not necessarily lead to correct diagnosis. From all "sure" and "absolutely sure" records, only 51.8% of the diagnosis were correct (Data in supplement **S5 – S6**).

D/ Comparison with the AI software

The in-house developed AI-based software perfectly matched the pattern interpretations of ATS/ERS guidelines (100%). Software response was 0.2 seconds, giving immediate and consistent interpretations. Moreover, it assigned a correct diagnostic category in 82% of the cases, which was highly superior to the average 44.6% accuracy of the pulmonologists ($p <0.0001$) (**Figure 2)**. It also proved to be highly sensitive in recognising COPD, neuromuscular disease, interstitial lung disease and healthy subjects. Concerning PPV, the software showed powerful results for the majority of the respiratory disease diagnoses (details in **Figure 3** and **Table 4**). Both sensitivity and positive predictive value of the AI-based algorithm were superior to expert-based diagnostic category allocation in each of the eight disease groups (**Figure 3**). AI lacked sensitivity for the OBD group, which was recouped by the very high PPV.

DISCUSSION

In this study, we explored the accuracy and the consistency between pulmonologists when interpreting PFT patterns and providing a preferred diagnostic category. PFT pattern interpretations matched the ATS/ERS guidelines in 74.4% of the cases with an inter-rater variability of 0.67, demonstrating that such a fundamental task is prone to mistakes and disagreements. PFTs combined with limited clinical information were difficult for pulmonologists as the only tool for reaching an accurate diagnostic category (accuracy of 44.6% and significant variability of kappa 0.35). However, our advanced AI-based software for the automated clinical reading of PFTs, perfectly interpreted (100%) PFT patterns and pointed to the correct diagnostic category in 82% of all cases. Consequently, it outperformed the pulmonologists in both tasks by 34% and 84% respectively, which demonstrates that individual pulmonologists do not sufficiently capture the information available in PFTs.

Facilitating clinical practice with decision support systems is not a new idea, and it has been shown that a majority (64%) of such systems do improve performances of individual clinicians.[22] Nowadays, we regularly use them to interpret electrocardiograms, to analyse mammogram irregularities or as reminders for drug prescription.[23, 24] Although automated analysis of PFTs had been evaluated before[25, 26], none has become a clinical reality. First, there is an obvious difficulty in reaching a preferred diagnosis without knowing the clinical context.[27, 28] Second, there is a lack of clear international diagnostic guidelines to label respiratory diseases based on PFTs with controversial and often arbitrary choice of cut-offs to label abnormality. It implies that not all pulmonologists are using the same interpretative strategies in their daily routine.[29, 30] For example, a typical conflict is often seen in the first interpretative step: should we take the lower limits of normal or fixed 0.7 cut-off for FEV_1/FVC ratio.[31] Undoubtedly, it will explain some of the differences between the interpretations of pulmonologists, but it also highlights a more general concern. Different

recommendations on which cut-offs to use will reclassify individual patients from healthy to diseased and vice versa, while in real life disease processes will present as a continuum around prefixed values. The strength of complete PFTs lies in the variety and multitude of tests in order to recognise disease-specific patterns, regardless of these fixed cut-off points.

Using the AI, we approached each disease as having a unique fingerprint on the PFT. As such, AI identifies subtle and defining characteristics that are challenging for humans to detect and incorporates them into a powerful discriminating diagnostic algorithm. In our case, AI takes complete input data and maps them into a high dimensional space. As a result of a large number of known disease cases, with known magnitudes and patterns between all input data, AI will construct most optimal hyperplanes which categorise new examples. Once presented with the data of a new patient, AI maps them into the same high dimensional space and predict to which category it belongs. Such a multi-dimensional approach exceeds human capabilities to observe the same data in terms of accuracy. Fundamentally, AI algorithm is not anymore dependent on the arbitrary cut-offs but a purely patient-data-driven knowledge system. In fact, with the increase in computing resources, modern AI algorithms have entirely moved away from rule-based systems and currently adopt a probabilistic approach. Our study confirms that a unique data-driven fingerprint of each disease often exists in the PFTs.

A fascinating characteristic of an AI-based software is its ability to improve over time by getting exposure to new and more difficult cases. In other words, the developed software may improve - as physicians do - by learning from mistakes and gaining experience. It is too ambitious to expect the software to be correct in 100% of the cases, as some respiratory diseases do not show characteristic lung function abnormalities. Particularly for early disease stages or combined complex disease processes, disease-specific characteristics may be hidden. As the current accuracy of the AI software is situated within the range of what clinical expert panels reached during the Belgian pulmonary lung function study[8], there is

probably little room for improvement. However, it also indicates that a computer can process all necessary information as effectively as a group of experts - not the individual - yet at a much higher speed and with hundred percent consistency for the same data input. The further usefulness of the AI software will be demonstrated if it decreases the time to final diagnosis, reduces the number of tests needed for a final diagnosis, and if by standardizing PFT interpretation, a number of misdiagnoses can be avoided.

Comparable with the human examiner marking his confidence on the Likert scale, AI expresses its certainty as a probability for a patient to belong to one of the disease categories. In the situations where AI made a wrong diagnostic suggestion, it should be mentioned that it never attributed a high probability to this diagnosis. More specifically, probability barely crossed 50% in two out of the nine mislabels and it was lower than 50% in the seven other cases. Surprisingly, the use of the COPD assessment test (CAT) for the quantification of symptoms in the BPFS study, did not contribute to further improving the accuracy of our AI software. It suggests that most respiratory diseases present with similar non-specific symptoms such as cough and dyspnoea. It is tempting to speculate that more input, such as more extensive history taking, and tests like exhaled nitric oxide, forced oscillometry and/or blood/radiological markers, could enhance its future potential. In particular, for diseases such as asthma which can perfectly present with a normal PFT, the added value of such tests when integrated into our AI-based software, is obvious.

A limitation of the current study is that we underestimated the accuracy of the pulmonologists by limiting the amount of clinical data to suggest a preferred diagnosis. In reality, a diagnosis is reached by a synergy of multiple factors, including expanded history, clinical examination, imaging and blood sampling. The real-life situation may therefore yield better outcomes. Additionally, the test sample we used may not entirely reflect the prevalence of diseases that pulmonologists confront in daily clinical practice. It is clear that we only explored the

maximum output that could be reached from PFTs and clinical information, representative of the first diagnostic encounter. Furthermore, we did not formally test the level of agreement within the ad-hoc expert panel to define the final diagnosis. Although the experts relied on all available test information, one may speculate that providing the AI interpretation would have favoured their initial agreement. A final limitation is that the risk of misinterpretation and misdiagnosis increases if tests are poorly performed.[32] However, sufficient quality of the tests is needed for both human and computer interpretations.

To conclude, our data indicates that interpretation of PFTs and the suggestion of primary respiratory disease diagnosis by pulmonologists is highly variable. The AI-based software has superior performance and may provide a powerful decision support tool for clinicians. The significance of such technology in improving clinical practice will drive real-life acceptance of the medical community.

.

Acknowledgments: We thank all the pulmonologists, the pulmonary function technicians, the patients and the hospitals who participated in the study for providing and analyzing data.

Author contributions: All authors critically revised the manuscript and approved the final version. All authors organised evaluation sessions in hospitals, examined patient files, and interpreted results. MT performed the data acquisition, analysis, interpretation as well as contributed to the study design and wrote the manuscript. ND contributed to data acquisition. WJ takes responsibility for the content of the manuscript, contributed to the study design and assisted in the data analysis, interpretation and writing of the manuscript.

Conflict of interest: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. No other disclosures related to this work were reported.

Support statement: This work was supported by the Vlaams Agentschap Innoveren & Ondernemen - VLAIO, Government body, 2016 - 2018. The funder had no role in study design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

The Pulmonary function study investigators

R. De Pauw, C. Depuydt, C. Haenebalcke, S. Muyldermans, V. Ringoet, D. Stevens (AZ Sint Jan, Brugge); S. Bayat, J. Benet, E. Catho, J. Claustre, A. Fedi, MA. Ferjani, R. Guzun, M. Isnard, S. Nicolas, T. Pierret, C. Pison, S. Rouches, B. Wuyam (CHU, Grenoble); JL. Corhay, J. Guiot, K. Ghysen, L. Renaud, A. Sibille (CHU, Liège); H. De La Barriere, C. Charpentier, S. Corhut, KA. Hamdan, M. Schlesser, G. Wirtz (CHU, Luxembourg); E. Alabadan, G. Birsen, PR. Burgel, A. Chohra, C. Hamard, B. Lemarié, MN. Lothe, C. Martin, AC. Sainte-Marie, L. Sebane (Cochin Hospital, Paris); Y. Berk, B. de Brouwer, R. Janssen, J. Kerkhoff,

A. Spaanderman, M. Stegers, A. Termeer, I. van Grimbergen, A. van Veen, L. van Ruitenbeek, L. Vermeer, R. Zaal, M. Zijlker (CW Hospital, Nijmegen); J. Aumann, K. Cuppens, D. Degraeve, K. Demuynck, B. Dieriks, K. Pat, L. Spaas, R. Van Puijenbroek, K. Weytjens, J. Wynants (Jessa Hospital, Hasselt); V. Adam, BJ. Berendes, E. Hardeman, P. Jordens, E. Munghen, K. Tournoy, P. Vercauter (OLV Ziekenhuis, Aalst); T. Alame, M.Bruyneel, M. Gabrovska, I.Muylle, V.Ninane, D. Rozen, P. Rummens. S. Van Den Broecke, (St. Pierre, Bruxelles); A. Froidure, S. Gohy, G.Liistro, T. Pieters, C. Pilette, F. Pirson (UCL, Bruxelles); H. Kerstjens, M. Van den Berge, N. Ten Hacken, M. Duiverman, D. Koster (UMC, Groningen); B. Vosse, L. Conemans, M. Maus, M. Bischoff, M. Rutten, D. Agterhuis, R. Sprooten (UMC, Maastricht); B. Beutel, A. Jerrentrup, A. Klemmer, C. Viniol, C. Vogelmeier (UMC, Marburg); H. Bode, C. Dooms, D. Gullentops, W. Janssens, K. Nackaerts, D. Rutens, E. Wauters, W. Wuyts (UZ Gasthuisberg, Leuven); E. Derom, S. Dobbelaere, S. Loof, G. Serry, B. Putman, L. Van Acker, Y. Vandeweygaerde (UZ, Gent); M. Criel, M. Daenen, R. Gubbelmans, S. Klerkx, E. Michiels, M. Thomeer, A. Vanhauwaert (ZOL Genk).

REFERENCE LIST

1. Crapo RO. Pulmonary-function testing. *New England Journal of Medicine* 1994: 331(1): 25-30.

2. Pellegrino R, Viegi G, Brusasco V, Crapo RO, Burgos F, Casaburi R, Coates A, van der Grinten CP, Gustafsson P, Hankinson J, Jensen R, Johnson DC, MacIntyre N, McKay R, Miller MR, Navajas D, Pedersen OF, Wanger J. Interpretative strategies for lung function tests. *Eur Respir J* 2005: 26(5): 948-968.

3. Reddel HK, Bateman ED, Becker A, Boulet LP, Cruz AA, Drazen JM, Haahtela T, Hurd SS, Inoue H, de Jongste JC, Lemanske RF, Jr., Levy ML, O'Byrne PM, Paggiaro P, Pedersen SE, Pizzichini E, Soto-Quiroz M, Szefler SJ, Wong GW, FitzGerald JM. A summary of the new GINA strategy: a roadmap to asthma control. *Eur Respir J* 2015: 46(3): 622-639.

4. Vogelmeier CF, Criner GJ, Martinez FJ, Anzueto A, Barnes PJ, Bourbeau J, Celli BR, Chen R, Decramer M, Fabbri LM, Frith P, Halpin DM, Lopez Varela MV, Nishimura M, Roche N, Rodriguez-Roisin R, Sin DD, Singh D, Stockley R, Vestbo J, Wedzicha JA, Agusti A. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 Report: GOLD Executive Summary. *Eur Respir J* 2017: 49(3).

5. Galie N, Humbert M, Vachiery JL, Gibbs S, Lang I, Torbicki A, Simonneau G, Peacock A, Vonk Noordegraaf A, Beghetti M, Ghofrani A, Gomez Sanchez MA, Hansmann G, Klepetko W, Lancellotti P, Matucci M, McDonagh T, Pierard LA, Trindade PT, Zompatori M, Hoeper M. 2015 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension: The Joint Task Force for the Diagnosis and Treatment of Pulmonary Hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS): Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC), International Society for Heart and Lung Transplantation (ISHLT). *Eur Respir J* 2015: 46(4): 903-975.

6. Martinez FJ, Chisholm A, Collard HR, Flaherty KR, Myers J, Raghu G, Walsh SL, White ES, Richeldi L. The diagnosis of idiopathic pulmonary fibrosis: current and future approaches. *Lancet Respir Med* 2017: 5(1): 61-71.

7. Topalovic M, Laval S, Aerts JM, Troosters T, Decramer M, Janssens W, Belgian Pulmonary Function Study i. Automated Interpretation of Pulmonary Function Tests in Adults with Respiratory Complaints. *Respiration* 2017: 93(3): 170-178.

8. Decramer M, Janssens W, Derom E, Joos G, Ninane V, Deman R, Van Renterghem D, Liistro G, Bogaerts K, Investigators BPFS. Contribution of four common pulmonary function tests to diagnosis of patients with respiratory symptoms: a prospective cohort study. *The Lancet Respiratory Medicine* 2013: 1(9): 705-713.

9. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017: 542(7639): 115-118.

10. Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, Hamzah H, Garcia-Franco R, San Yeo IY, Lee SY, Wong EYM, Sabanayagam C, Baskaran M, Ibrahim F, Tan NC, Finkelstein EA, Lamoureux EL, Wong IY, Bressler NM, Sivaprasad S, Varma R, Jonas JB, He MG, Cheng CY, Cheung GCM, Aung T, Hsu W, Lee ML, Wong TY. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA* 2017: 318(22): 2211-2223.

11. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, van der Laak J, the CC, Hermsen M, Manson QF, Balkenhol M, Geessink O, Stathonikos N, van Dijk MC, Bult P, Beca F, Beck AH, Wang D, Khosla A, Gargeya R, Irshad H, Zhong A, Dou Q, Li Q, Chen H, Lin HJ, Heng PA, Hass C, Bruni E, Wong Q, Halici U, Oner MU, Cetin-Atalay R,

Berseth M, Khvatkov V, Vylegzhanin A, Kraus O, Shaban M, Rajpoot N, Awan R, Sirinukunwattana K, Qaiser T, Tsang YW, Tellez D, Annuscheit J, Hufnagl P, Valkonen M, Kartasalo K, Latonen L, Ruusuvuori P, Liimatainen K, Albarqouni S, Mungal B, George A, Demirci S, Navab N, Watanabe S, Seno S, Takenaka Y, Matsuda H, Ahmady Phoulady H, Kovalev V, Kalinovsky A, Liauchuk V, Bueno G, Fernandez-Carrobles MM, Serrano I, Deniz O, Racoceanu D, Venancio R. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA* 2017: 318(22): 2199-2210.

12. Thrall JH, Li X, Li Q, Cruz C, Do S, Dreyer K, Brink J. Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success. *J Am Coll Radiol* 2018: 15(3 Pt B): 504-508.

13. Armstrong S. The computer will assess you now. *BMJ* 2016: 355: i5680.

14. Fridsma DB. Health informatics: a required skill for 21st century clinicians. *BMJ* 2018: 362: k3043.

15. The L. Artificial intelligence in health care: within touching distance. *Lancet* 2018: 390(10114): 2739.

16. Culver BH, Graham BL, Coates AL, Wanger J, Berry CE, Clarke PK, Hallstrand TS, Hankinson JL, Kaminsky DA, MacIntyre NR, McCormack MC, Rosenfeld M, Stanojevic S, Weiner DJ, Laboratories ATSCoPSfPF. Recommendations for a Standardized Pulmonary Function Report. An Official American Thoracic Society Technical Statement. *Am J Respir Crit Care Med* 2017: 196(11): 1463-1472.

17. Topalovic M, Das N, Troosters T, Decramer M, Janssens W. Late Breaking Abstract - Applying artificial intelligence on pulmonary function tests improves the diagnostic accuracy. *European Respiratory Journal* 2017: 50(suppl 61).

18. Miller MR, Crapo R, Hankinson J, Brusasco V, Burgos F, Casaburi R, Coates A, Enright P, van der Grinten CP, Gustafsson P, Jensen R, Johnson DC, MacIntyre N, McKay R, Navajas D, Pedersen OF, Pellegrino R, Viegi G, Wanger J, Force AET. General considerations for lung function testing. *Eur Respir J* 2005: 26(1): 153-161.

19. Quanjer PH, Stanojevic S, Cole TJ, Baur X, Hall GL, Culver BH, Enright PL, Hankinson JL, Ip MS, Zheng J, Stocks J, Initiative ERSGLF. Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations. *Eur Respir J* 2012: 40(6): 1324-1343.

20. Quanjer PH, Tammeling G, Cotes J, Pedersen O, Peslin R, Yernault J. Lung volumes and forced ventilatory flows. *European Respiratory Journal* 1993: 6(Suppl 16): 5-40.

21. Stanojevic S, Graham BL, Cooper BG, Thompson BR, Carter KW, Francis RW, Hall GL, Global Lung Function Initiative Twg, Global Lung Function Initiative T. Official ERS technical standards: Global Lung Function Initiative reference values for the carbon monoxide transfer factor for Caucasians. *Eur Respir J* 2017: 50(3).

22. Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, Sam J, Haynes RB. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA* 2005: 293(10): 1223-1238.

23. Filippi A, Sabatini A, Badioli L, Samani F, Mazzaglia G, Catapano A, Cricelli C. Effects of an automated electronic reminder in changing the antiplatelet drug-prescribing behavior among Italian general practitioners in diabetic patients: an intervention trial. *Diabetes Care* 2003: 26(5): 1497-1500.

24. Willems JL, Abreu-Lima C, Arnaud P, van Bemmel JH, Brohet C, Degani R, Denis B, Gehring J, Graham I, van Herpen G. The diagnostic performance of computer programs for the

interpretation of electrocardiograms. *New England Journal of Medicine* 1991: 325(25): 1767-1773.

25. Hankinson JL. Automated pulmonary function testing: interpretation and standardization. *Ann Biomed Eng* 1981: 9(5-6): 633-643.

26. Krumpe P, Weigt G, Martinez N, Marcum R, Cummiskey JM. Computerized rapid analysis of pulmonary function test: use of a least mean squares correlation for interpretation of data. *Comput Biol Med* 1982: 12(4): 295-307.

27. Berry CE, Wise RA. Interpretation of pulmonary function test: issues and controversies. *Clin Rev Allergy Immunol* 2009: 37(3): 173-180.

28. Enright P. Flawed interpretative strategies for lung function tests harm patients. *European Respiratory Journal* 2006: 27(6): 1322-1323.

29. Miller MR, Quanjer PH, Swanney MP, Ruppel G, Enright PL. Interpreting lung function data using 80% predicted and fixed thresholds misclassifies more than 20% of patients. *Chest* 2011: 139(1): 52-59.

30. Visentin E, Nieri D, Vagaggini B, Peruzzi E, Paggiaro P. An observation of prescription behaviors and adherence to guidelines in patients with COPD: real world data from October 2012 to September 2014. *Curr Med Res Opin* 2016: 32(9): 1493-1502.

31. Quanjer PH, Enright PL, Miller MR, Stocks J, Ruppel G, Swanney MP, Crapo RO, Pedersen OF, Falaschetti E, Schouten JP, Jensen RL. The need to change the method for defining mild airway obstruction. *Eur Respir J* 2011: 37(3): 720-722.

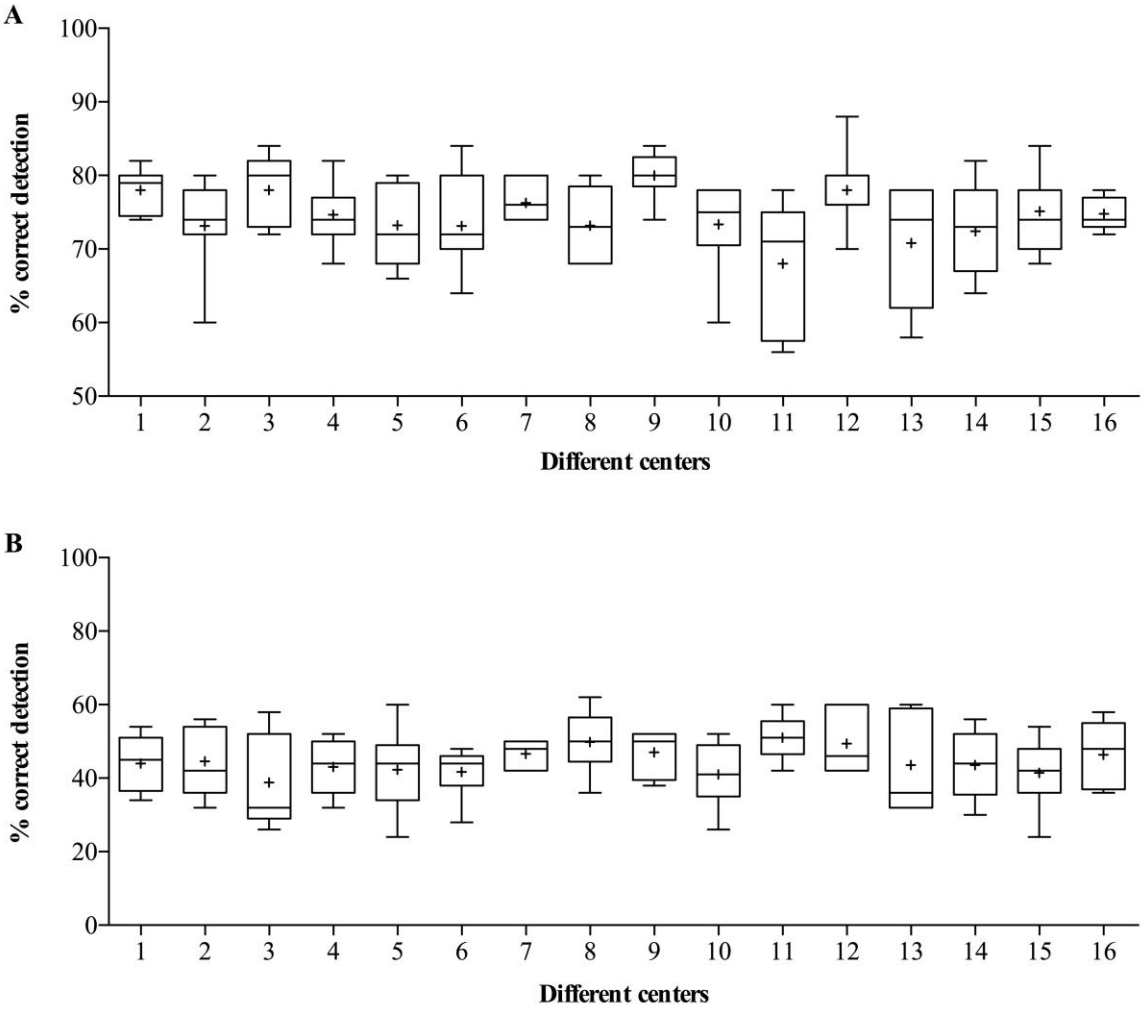
32. Leuppi JD, Miedinger D, Chhajed PN, Buess C, Schafroth S, Bucher HC, Tamm M. Quality of spirometry in primary care for case finding of airway obstruction in smokers. *Respiration* 2010: 79(6): 469-474.

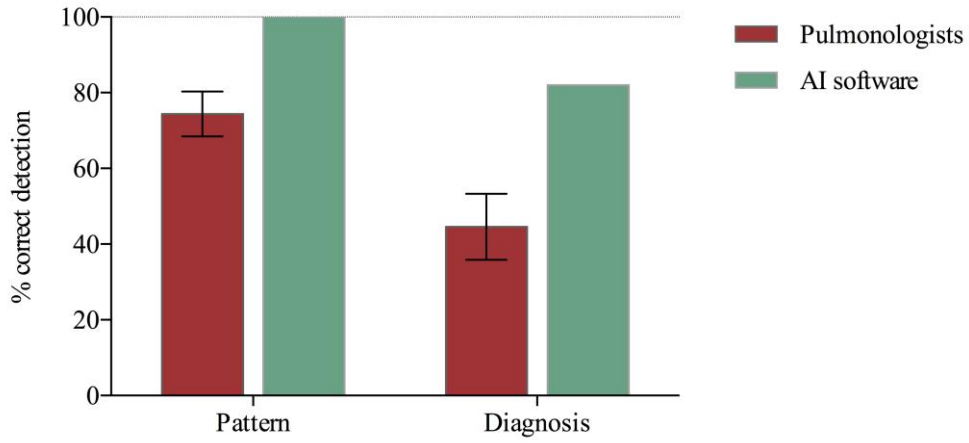
FIGURE LEGENDS

Figure 1: Comparison of correct detections per each centre. **Panel A**/ PFT pattern interpretation. No significant difference between centres (p=0.06). **Panel B**/ Preferred diagnostic category. No significant difference between centres (p=0.44). Data anonymised. Data presented: line at the median (IQR), + at the mean, range.

Figure 2: Comparison of results obtained by pulmonologists (red bars) versus results achieved by AI- software (green bars). Correct detections are significantly ($p<0.0001$) higher for AI-software (improvement of 34% for PFT pattern interpretation and 84% for diagnostic category detection).

Figure 3: Performance of pulmonologists in comparison with the AI software for each disease category. Left bars are values obtained when software analysed study population; right bars are values obtained by pulmonologists' evaluation. **Panel A/** Sensitivity (=True positive / (True positive + False Negative) shows how many relevant subjects (from a specific group) were correctly identified. **Panel B/** Positive predictive value (=True positive / (True positive + False Positive) shows how many labelled subjects rightly belonged to the specific group. COPD = Chronic Obstructive Pulmonary Disease, ILD = Interstitial Lung Disease, NMD = Neuromuscular Disease, OBD = Other Obstructive Diseases, PVD = Pulmonary Vascular Disease, TD = Thoracic Deformity.


A
100
90
80
70
60
50
% correct detection
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Different centers
B
100
80
60
40
20
0
% correct detection
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Different centers


100
80
60
40
20
0
% correct detection
Pattern
Diagnosis
Pulmonologists
AI software

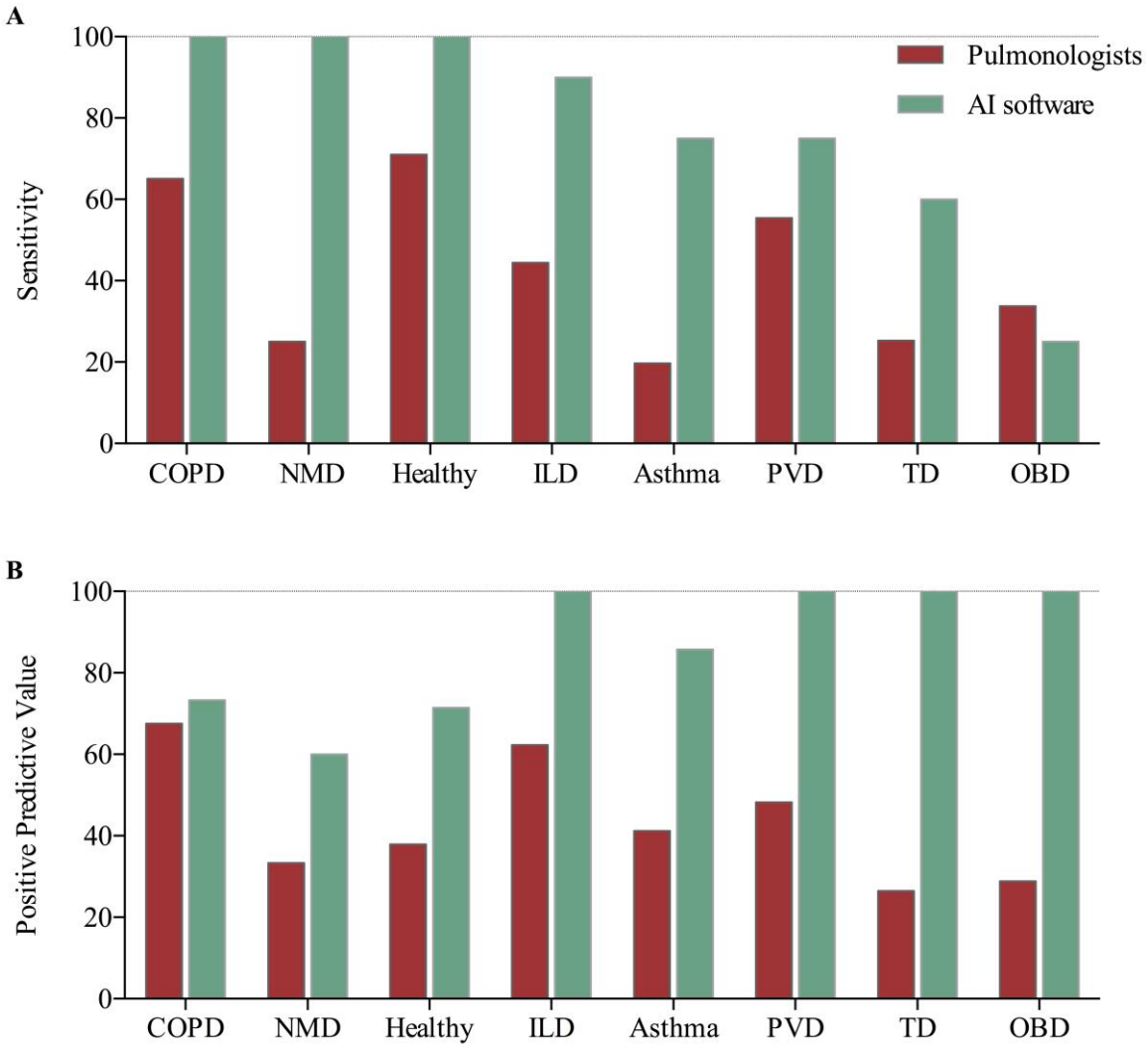

A
100
80
60
40
20
0
Sensitivity
Pulmonologists
AI software
COPD
NMD
Healthy
ILD
Asthma
PVD
TD
OBD
B
100
80
60
40
20
0
Positive Predictive Value
COPD
NMD
Healthy
ILD
Asthma
PVD
TD
OBD

Table 1: Population characteristics of the 50 subjects whose lung function was evaluated in the study

	Asthma	COPD	OBD	NMD	TD	ILD	PVD	Healthy
Subjects, n	8	11	4	3	5	10	4	5
Sex, M/F	5/3	8/3	3/1	2/1	4/1	6/4	3/1	3/2
Age, years	57 (27 - 70)	64 (38 - 77)	53 (34 - 77)	65 (48 - 72)	60 (52 - 68)	70 (51 - 83)	80 (62 - 81)	64 (38 – 74)
FEV_1, z-score.	-0.57 (-2.70 to 0.73)	-1.41 (-3.95 to 0.41)	-2.97 (-4.05 to -1.39)	-2.47 (-2.87 to -1.97)	-2.76 (-2.94 to -1.77)	-0.65 (-2.74 to 1.01)	0.17 (-2.23 to 0.78)	0.24 (-0.01 to 1.63)
FVC, z-score	-0.41 (-1.86 to 2.00)	0.70 (-2.48 to 2.07)	-2.51 (-4.37 to -0.48)	-2.58 (-2.85 to -1.93)	-2.68 (-2.79 to -2.30)	-0.97 (-3.42 to 0.79)	0.66 (-1.83 to 1.59)	0.11 (-0.06 to 1.22)
FEV_1/FVC, z-score	-1.01 (-2.79 to 0.29)	-2.54 (-4.86 to -1.54)	-2.51 (-4.37 to -0.48)	-0.41 (-0.60 to 0.07)	-0.83 (-1.41 to 1.47)	0.85 (-0.25 to 2.05)	-0.90 (-1.10 to -0.53)	0.32 (-0.26 to 0.50)
TLC, z-score	0.01 (-1.04 to 2.39)	1.55 (-1.49 to 2.80)	0.17 (-0.74 to 1.20)	-2.23 (-3.01 to -2.17)	-2.98 (-5.05 to -1.10)	-2.54 (-4.96 to -1.00)	-0.29 (-1.53 to 0.11)	-0.13 (-0.43 to 1.50)
RV, z-score.	-0.02 (-2.81 to 4.49)	1.10 (-1.59 to 6.24)	2.24 (1.38 to 3.22)	-0.95 (-1.08 to -0.19)	-1.50 (-2.98 to 1.67)	-2.45 (-4.20 to -1.35)	-0.79 (-2.34 to 0.16)	-0.99 (-2.42 to 3.14)
DL_{CO}, z-score	-0.84 (-1.96 to 1.25)	-2.77 (-4.39 to -0.54)	-1.89 (-3.98 to -0.67)	-2.08 (-2.30 to -1.74)	-2.44 (-4.77 to -1.98)	-2.91 (-4.30 to -0.06)	-2.80 (-4.17 to -2.33)	-0.67 (-2.37 to -0.29)
K_{CO}, z-score	0.09 (-0.93 to 1.48)	-2.05 (-2.93 to -0.27)	-0.17 (-1.94 to 1.95)	0.53 (0.48 to 0.55)	0.18 (-1.86 to 1.73)	-1.09 (-2.04 to 1.27)	-2.02 (-3.53 to -1.17)	-0.32 (-1.34 to -0.07)

Definition of abbreviations: FEV_1 = forced expiratory volume in one second; F = Female; FVC = forced vital capacity; DL_{CO} = diffusing capacity for carbon monoxide; K_{CO} = transfer coefficient for carbon monoxide; M= Male; TLC = total lung capacity; NMD = neuromuscular disease; ILD = interstitial lung diseases; PVD = pulmonary vascular diseases; OBD = other obstructive diseases; TD = Thoracic deformity/ Pleural diseases; COPD = chronic obstructive pulmonary disease. Values are median and range.

Table 2: Confusion matrix with counts of all correctly and incorrectly labelled subjects per PFT pattern

		Pulmonologist Pattern					
		Obstructive	**Restrictive**	**Normal**	**Mixed**	**Total**	**N Subjects**
Reference Pattern	**Obstructive**	**1636**	196	180	112	2124	18
	Restrictive	34	**883**	14	249	1180	10
	Normal	285	424	**1872**	15	2596	22
	Mixed	0	0	0	**0**	0	0
	Total	1955	1503	2066	376	5900	
	Avg. N Sub.	17	13	17	3		50
	Specificity	92%	87%	94%	.		
	Sensitivity	77%	75%	72%	.		
	PPV	84%	59%	91%	.		
	NPV	88%	93%	81%	.		

Rows show true reference PFT patterns, while columns show patterns labelled by pulmonologists. There are 4391 (=74.4%) correctly given interpretations (true positive in bold.). PPV, positive predictive value; NPV, negative predictive value; Avg. N Sub., averaged N subjects for each pattern given by each pulmonologist.

Table 3: Confusion matrix with counts of all correctly and incorrectly labelled subjects per each diagnostic category

		Pulmonologist Diagnosis										
		Asthma	**COPD**	**OBD**	**NMD**	**TD**	**ILD**	**PVD**	**Healthy**	**Other**	**Total**	**N Subjects**
Reference Diagnosis	**Asthma**	**189**	82	141	23	49	4	5	395	72	960	8
	COPD	157	**859**	154	4	6	28	49	22	41	1320	11
	OBD	77	139	**162**	13	15	5	6	45	18	480	4
	NMD	1	2	7	**90**	156	70	3	4	27	360	3
	TD	10	103	56	68	**152**	133	7	15	56	600	5
	ILD	2	9	5	58	168	**533**	167	205	53	1200	10
	PVD	2	55	27	8	18	75	**266**	11	18	480	4
	Healthy	21	24	10	6	9	7	49	**426**	48	600	5
	Other	0	0	0	0	0	0	0	0	**0**	0	0
	Total	459	1273	562	270	573	855	552	1123	333	6000	
	Avg. N Sub.	3.8	10.6	4.7	2.3	4.8	7.1	4.6	9.4	2.8		50
	Specificity	90%	81%	86%	93%	86%	87%	89%	76%	.		
	Sensitivity	20%	65%	34%	25%	25%	44%	55%	71%	0%		
	PPV	41%	67%	29%	33%	27%	62%	48%	38%	0%		
	NPV	76%	80%	89%	91%	85%	76%	92%	93%	.		

Rows show true reference diagnostic category, while columns show diagnosis labelled by pulmonologists. There are 2667 (=44.6%) correctly suggested diagnosis (true positive in bold.). PPV, positive predictive value; NPV, negative predictive value; Avg. N Sub., averaged N subjects for each diagnostic category given by each pulmonologist.

Table 4: Confusion matrix with counts of all correctly and incorrectly labelled subjects by AI software per each diagnostic category

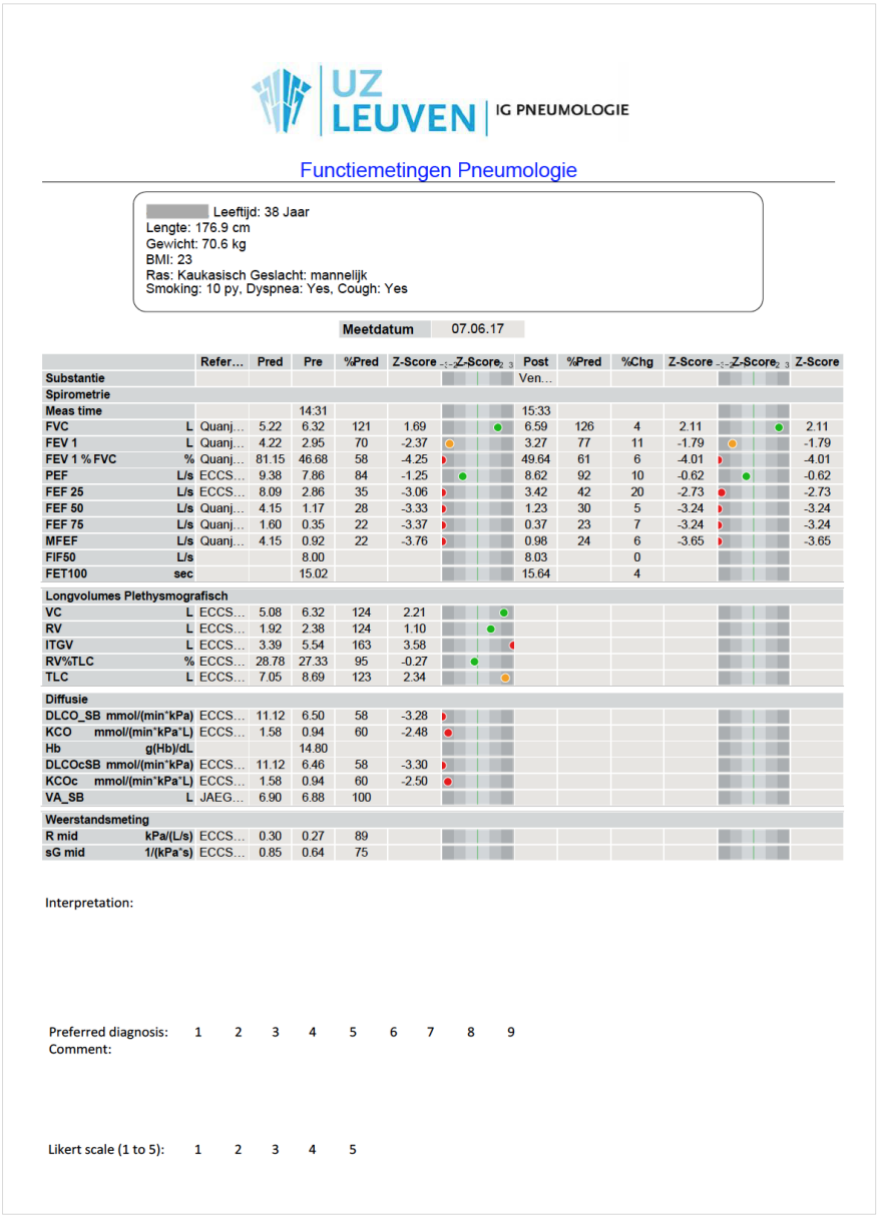
		AI software Diagnosis									
		Asthma	**COPD**	**OBD**	**NMD**	**TD**	**ILD**	**PVD**	**Healthy**	**Other**	**N Subjects**
Reference Diagnosis	**Asthma**	**6**	1	0	0	0	0	0	1	0	8
	COPD	0	**11**	0	0	0	0	0	0	0	11
	OBD	1	2	**1**	0	0	0	0	0	0	4
	NMD	0	0	0	**3**	0	0	0	0	0	3
	TD	0	0	0	2	**3**	0	0	0	0	5
	ILD	0	0	0	0	0	**9**	0	1	0	10
	PVD	0	1	0	0	0	0	**3**	0	0	4
	Healthy	0	0	0	0	0	0	0	**5**	0	5
	Other	0	0	0	0	0	0	0	0	**0**	0
	Total	7	15	1	5	3	9	3	7	0	50
	Specificity	97%	88%	100%	95%	100%	100%	100%	95%	.	
	Sensitivity	75%	100%	25%	100%	60%	90%	75%	100%	.	
	PPV	86%	73%	100%	60%	100%	100%	100%	71%	.	
	NPV	95%	100%	93%	100%	95%	97%	97%	100%	.	

Rows show true reference diagnostic category, while columns show diagnosis labelled by AI software. There are 41 (=82%) correctly suggested diagnosis (true positive in bold.). PPV, positive predictive value; NPV, negative predictive value;

SUPPLEMENTARY MATERIAL

Artificial intelligence outperforms pulmonologists in the interpretation of pulmonary function tests

Figure S1a: Example of a first page of the patient file with clinical information, complete pulmonary function tests and area for the evaluation. Example in Dutch language.



UZ LEUVEN | IG PNEUMOLOGIE

Functiemetingen Pneumologie

[redacted] Leeftijd: 38 Jaar
Lengte: 176.9 cm
Gewicht: 70.6 kg
BMI: 23
Ras: Kaukasisch Geslacht: mannelijk
Smoking: 10 py, Dyspnea: Yes, Cough: Yes

Meetdatum 07.06.17

		Refer...	Pred	Pre	%Pred	Z-Score	Z-Score	Post	%Pred	%Chg	Z-Score	Z-Score	Z-Score
Substantie								Ven...					
Spirometrie													
Meas time				14:31				15:33					
FVC	L	Quanj...	5.22	6.32	121	1.69		6.59	126	4	2.11		2.11
FEV 1	L	Quanj...	4.22	2.95	70	-2.37		3.27	77	11	-1.79		-1.79
FEV 1 % FVC	%	Quanj...	81.15	46.68	58	-4.25		49.64	61	6	-4.01		-4.01
PEF	L/s	ECCS...	9.38	7.86	84	-1.25		8.62	92	10	-0.62		-0.62
FEF 25	L/s	ECCS...	8.09	2.86	35	-3.06		3.42	42	20	-2.73		-2.73
FEF 50	L/s	Quanj...	4.15	1.17	28	-3.33		1.23	30	5	-3.24		-3.24
FEF 75	L/s	Quanj...	1.60	0.35	22	-3.37		0.37	23	7	-3.24		-3.24
MFEF	L/s	Quanj...	4.15	0.92	22	-3.76		0.98	24	6	-3.65		-3.65
FIF50	L/s			8.00				8.03		0			
FET100	sec			15.02				15.64		4			
Longvolumes Plethysmografisch													
VC	L	ECCS...	5.08	6.32	124	2.21							
RV	L	ECCS...	1.92	2.38	124	1.10							
ITGV	L	ECCS...	3.39	5.54	163	3.58							
RV%TLC	%	ECCS...	28.78	27.33	95	-0.27							
TLC	L	ECCS...	7.05	8.69	123	2.34							
Diffusie													
DLCO_SB	mmol/(min*kPa)	ECCS...	11.12	6.50	58	-3.28							
KCO	mmol/(min*kPa*L)	ECCS...	1.58	0.94	60	-2.48							
Hb	g(Hb)/dL			14.80									
DLCOcSB	mmol/(min*kPa)	ECCS...	11.12	6.46	58	-3.30							
KCOc	mmol/(min*kPa*L)	ECCS...	1.58	0.94	60	-2.50							
VA_SB	L	JAEG...	6.90	6.88	100								
Weerstandsmeting													
R mid	kPa/(L/s)	ECCS...	0.30	0.27	89								
sG mid	1/(kPa*s)	ECCS...	0.85	0.64	75								

Interpretation:

Preferred diagnosis: 1 2 3 4 5 6 7 8 9
Comment:

Likert scale (1 to 5): 1 2 3 4 5

Figure S1b: Example of a second page of the patient file with all curves from the tests available.

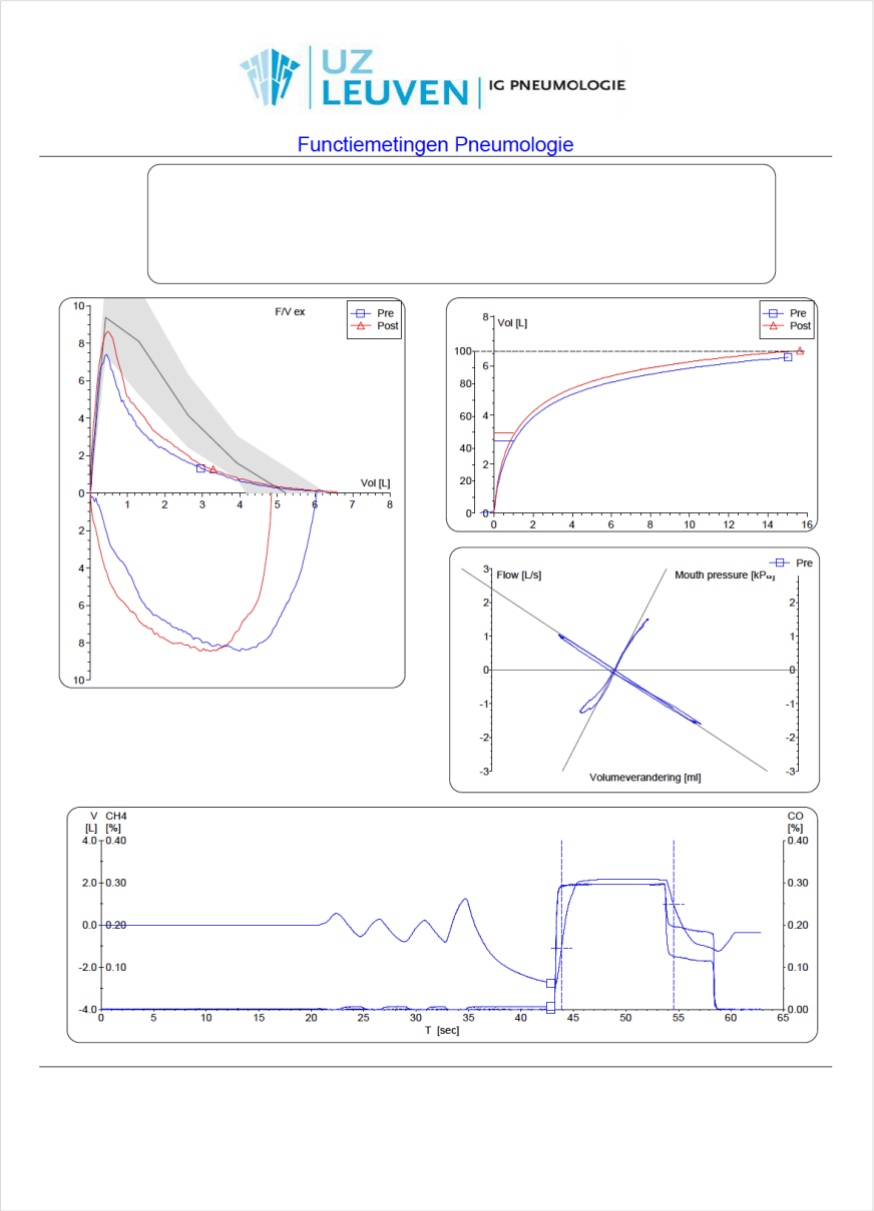


Figure S2: Protocol used to evaluate each patient case.

Lung function discussion

AIM 1:

Interpretation of the lung function tests.

Give description according to your preferences. Ideas:

- Pattern: obstructive, restrictive, mixed, normal
- Severity: mild, moderate, severe, very severe
- Special characteristics that you would highlight if you would make the protocol

AIM 2:

Give preferred diagnosis with the Likert scale (1 to 5) for the confidence in the decision

Likert scale:
1 point : absolutely not sure
2 points: not sure
3 points: some doubt
4 points: sure
5 points: absolutely sure

List of Diseases

1. **Asthma**
2. **COPD**
3. **Other obstructive disease** (OBD) [including: cystic fibrosis, bronchiectasis, bronchiolitis]
4. **Neuromuscular disease** (NMD) [including: paralysis of the diaphragm, poliomyelitis, myopathy]
5. **Thoracic deformity / Pleural disease** (TD) [including: pneumectomy, lobectomy, chest wall problems, kyphoscoliosis]
6. **Interstitial lung disease** (ILD) [including idiopathic pulmonary fibrosis, nonspecific interstitial pneumonitis and sarcoidosis]
7. **Pulmonary vascular disease** (PVD) [including pulmonary hypertension, embolism and vasculitis]
8. **Normal lung function** [Healthy]
9. **Other** (describe between brackets)

Figure S3: Report generated using in-house developed software. Probabilities for each diagnostic category summarized in the example.

Interpretation of lung function tests

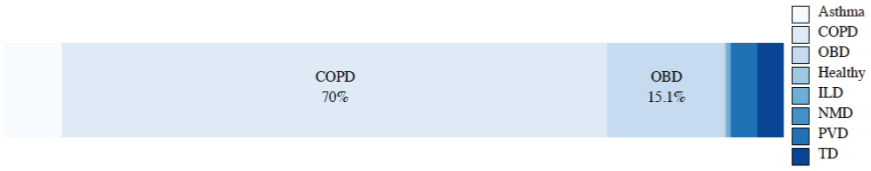
[redacted]

Patient with ID number: [redacted]
Analized on: [redacted]

This is a test sample

Protocol (based on interpretation rules):
Moderate obstructive lung function. Borderline significant reversibility after postbronchodilator test. Change in FEV1 of 320ml or 10.8%, and FVC of 270ml or 4.3%.
Normal airway resistance.
Hyperinflation. No signs of airtrapping.
Moderate reduction of diffusion capacity.

Disease probability (based on Artificial Intelligence):



Suggested diagnosis:
Lung function suggests that patient has: COPD.

Further suggestions:
Perform HRCT of the thorax.

contact: marko.topalovic@uzleuven.be

Figure S4: Comparison of diagnostic accuracy per disease averaged for all individual pulmonologists. **A/** Sensitivity (=True positive / (True positive + False Negative) shows how many relevant subjects (from specific group) were correctly identified. **B/** Positive predictive value (=True positive / (True positive + False Positive) shows how many labelled subjects rightly belonged to the specific group. COPD = Chronic Obstructive Pulmonary Disease, ILD = Interstitial Lung Disease, NMD = Neuromuscular Disease, OBD = Other Obstructive Diseases, PVD = Pulmonary Vascular Disease, TD = Thoracic Deformity.

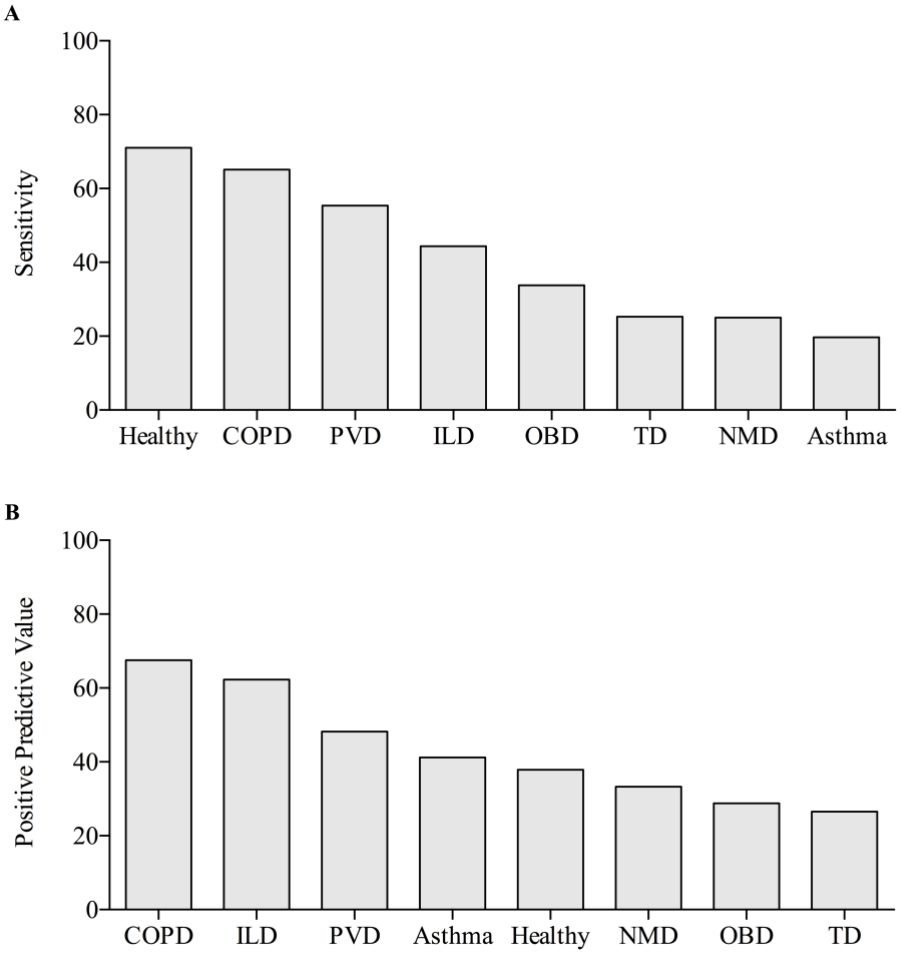


Figure S5: Distribution of Likert scores, with 3 (“some doubt”) and 4 (“sure”) being most common confidence answer.

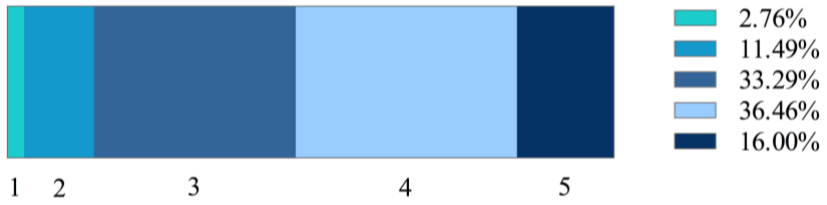


Figure S6: Comparison of Likert scores when decision was correct (median (IQR)= 4 (3-4)) versus when decision was not correct (=3 (3-4)).